Using Transfer Learning for Musical Instrument Classification

Alex L. Chandler
University of Texas at Austin Undergraduate
alex.chandler@utexas.edu

Abstract

For the past 300 years, luthiers have been responsible for not just making instruments, but identifying their maker, origin, and worth. Only in recent years has anyone tried to use machine learning for musical instrument classification. In this paper, I test the capabilities of pre-trained convolutional neural networks on the classification of musical instruments. I focus this paper on using AlexNet to predict the type of instrument (Violin, Viola, or Cello), country made of the instrument, and view of an instrument (front or back). I fine-tune AlexNet[1] and other convolutional neural networks on a database of 10,000 instruments. After fine-tuning, AlexNet yields a 99.8% accuracy for predicting the view of an instrument (front or back), 93% accuracy for predicting the type of instrument, and 65% for predicting country made on a balanced test dataset. Through a class saliency map, I show the parts of an image that are most important for instrument classification. The class saliency maps indicate that the neck, fingerboard, center bout, and body are all important components of classifying an instrument.

1 Introduction

Convolutional Neural Networks rely on large datasets to classify images with a high accuracy. The more data, the better. Since a collection of 10,000 instruments is not enough to train a deep neural network for a complex classification problem, I am modifying existing pre-trained deep neural networks. Pre-trained networks contain earlier layers that are good at detecting important image features. They learn much faster than a network with randomly initialized weights, and they generally require less data to yield high accuracy results. AlexNet is the name of a convolutional neural network that, although now outperformed by transformers, significantly outperformed competing models in 2012[2]. The AlexNet that I will be using for this project is accessible through PyTorch and was originally trained on over a million images from the ImageNet database.

2 The Dataset

The dataset I am using is from an extensive web scraping project I did back in 2020. The dataset includes data on 15,000 instruments. However, only about 8,000 instruments are without missing values. That dataset includes photos of the front of the instrument, back of

the instrument, instrument type, instrument maker, year-built, country made, city made, and sale data. For the purpose of this paper, I will be using photos of the instrument, country made, and instrument type.

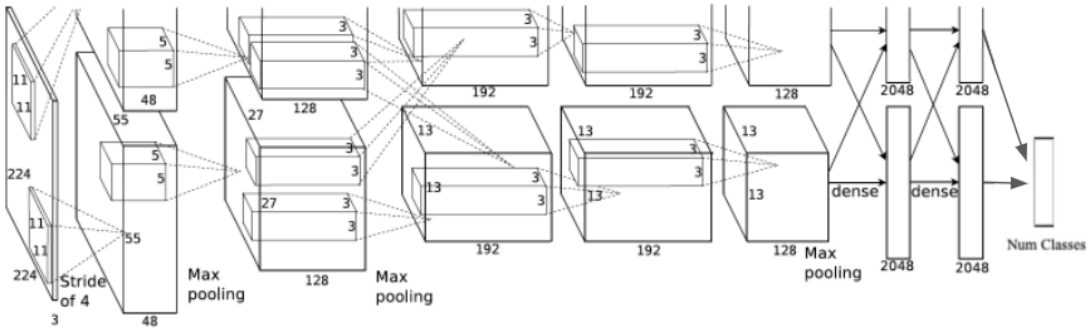
	image_top	maker_name	country_made	city_made	year_made	instrument_type	year_sold	sale_price_USD	image_back
0	[[[255, 255, 255], [255, 255, 255], [255, 255,...	Robert Clemens	United States	Chicago	NaN	Violin	2005.0	3738.0	[[[183, 183, 183], [183, 183, 183], [183, 183,...
1	[[[254, 254, 254], [254, 254, 254], [254, 254,...	John Frederick Lott II	United Kingdom	London	1840.0	Violin	NaN	NaN	[[[255, 255, 255], [255, 255, 255], [255, 255,...
2	[[[239, 239, 239], [239, 239, 239], [239, 239,...	John Frederick Lott II	United Kingdom	London	1840.0	Violin	NaN	NaN	[[[233, 233, 233], [233, 233, 233], [233, 233,...
3	[[[244, 244, 244], [244, 244, 244], [244, 244,...	John Frederick Lott II	United Kingdom	London	1850.0	Violin	NaN	NaN	[[[248, 248, 248], [248, 248, 248], [248, 248,...

3 Data Processing

The images are cropped and transposed from their original dimensions to (3,224,224). Padding is added on both the left and right sides so that each image is not stretched or squeezed. It is important to either maintain the original shape of the instrument or modify all images by the same height to width ratio, as the shape of an instrument is associated with the country where the instrument was built. Many rows did not include data on what country the instrument was made, but rather the city or region from where the instrument was made. To convert mixed locations into countries, I used the GeoPy library, which was able to convert over 90% of locations into countries.

4 Model Architecture

AlexNet is a deep convolutional neural network developed by Alex Krizhevsky (not me, unlike the name suggests) in 2012. It achieved unparalleled accuracy on predicting 1000 classes from the ImageNet database. Below is the standard AlexNet Architecture. For my initial instrument type, view, and country prediction models, I replace the final linear layer with a new linear layer with the number of output features equal to the desired number of classes.

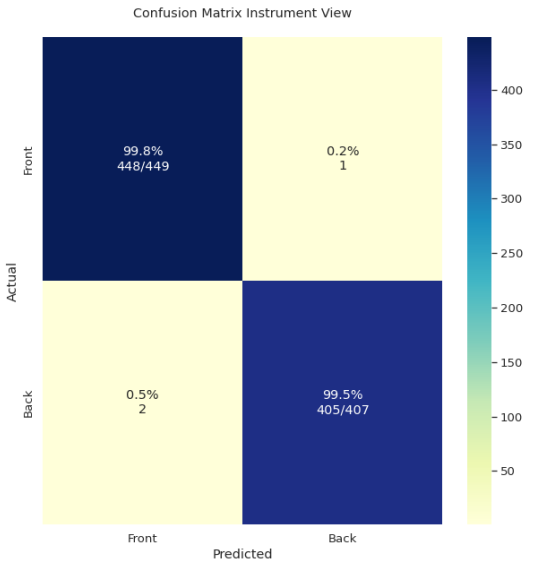


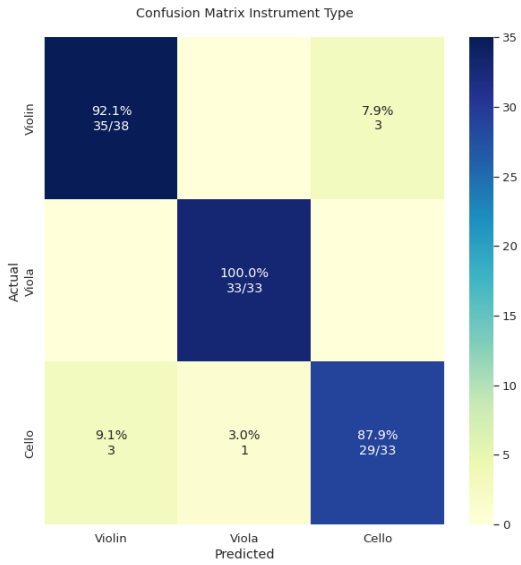
Evaluation

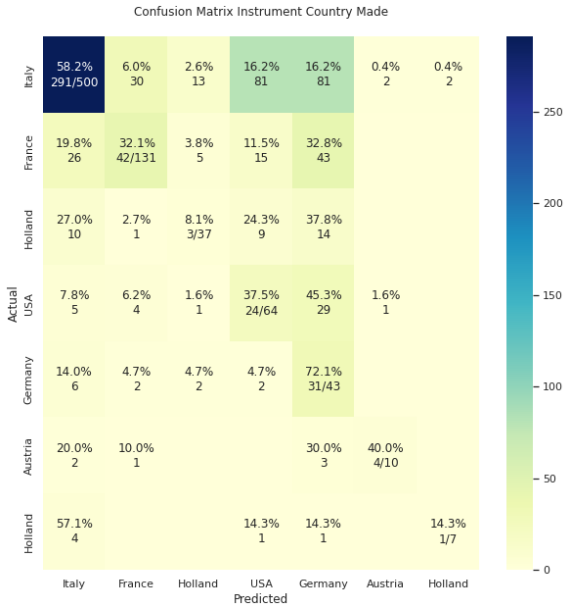
Fine-Tuned AlexNet with Modified Last Linear Layer Performance:

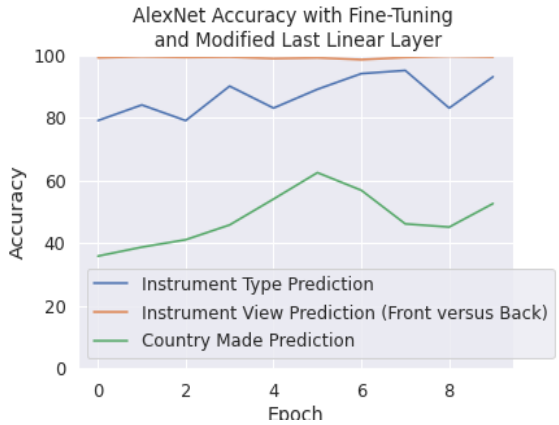
The confusion matrices below show the performance of pre-trained AlexNet with replacing the last linear layer fine-tuned over the instrument dataset. The instrument view classifier and instrument type classifier show excellent accuracy results in the confusion matrices below,

while the base country made classification model struggles to predict instruments from all locations with exception to predicting Italian and German instruments. As discussed in later computational experiments, I am able to increase model performance by adding an additional linear layer and instead of keeping the final linear of AlexNet.









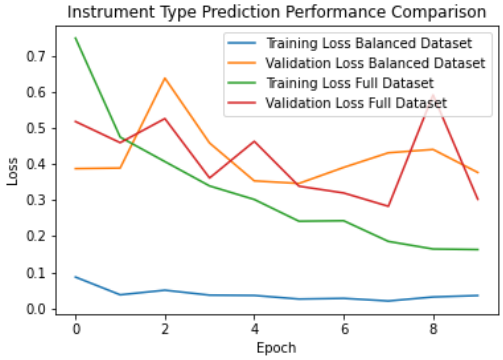
Computational Experiment 1: Balanced Dataset versus Weighted Loss

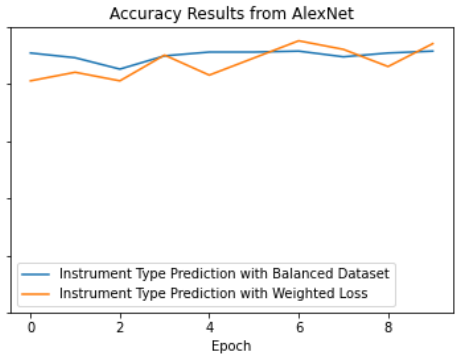
Question: How should I handle imbalanced class data?
The instrument dataset is extremely unbalanced. 86% of all instruments are violins, with the remaining 14% consisting mainly of violas and cellos. With a normal loss function and an unbalanced dataset, our model will struggle to learn and identify rare classes. Improvement across epochs may plateau if the model is not incentivised to predict less common classes. Two different solutions are discussed below.
Option 1:
Balance the dataset by sampling an equal number of violins, violas and cellos.

Option 2:
Have a different weight for each class (with larger weights for rare classes) and then normalize by the sum of the weights.

Hypothesis:
Option 2 will work slightly better, as the model will have more data to train on.

Results:





Using the full dataset with weighted loss (option 2) led to a significantly lower training loss and a slightly lower validation loss than balancing the dataset by removing examples from common classes. Similarly, training on the full dataset led to higher accuracy in later epochs; however, option 1 did surprisingly well considering the balanced dataset had only 2000 instruments compared to 8400 instruments in the full dataset. Of course, more testing would need to be done to confirm the difference in performance.

Computational Experiment 2: Image Normalization

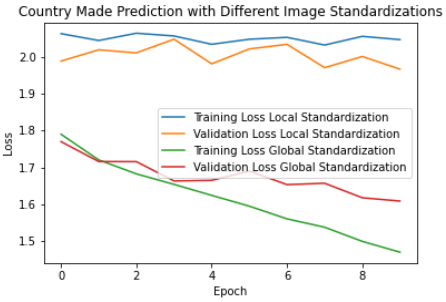
Question:
What types of Image Normalization and Standardization lead to the best performance?

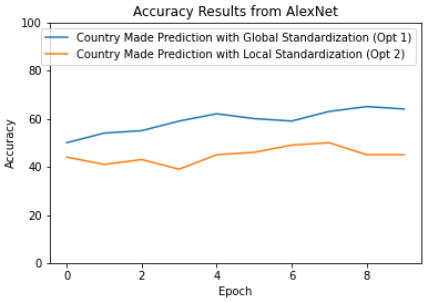
Option 1: Global Standardization
Calculate the standard deviation and mean across all images per RGB channel, then use these values to standardize the pixel values for each image.

Option 2: Local Standardization
Calculate the standard deviation and mean of one particular image per RGB channel, then use these statistics to standardize the pixels of the same particular image for each channel.

Hypothesis:
I believe that option 1 (Global Standardization) will lead to better higher performance for predicting the country made of an instrument. One of the key signatures to identify an instrument is its color. German makers generally preferred a brown varnish, while French makers tended to make instruments that were red, often called “French red". By standardizing color locally without considering the average color of other instruments, each processed image may lose distinctive color characteristics that may have helped the model make predictions.

Results:

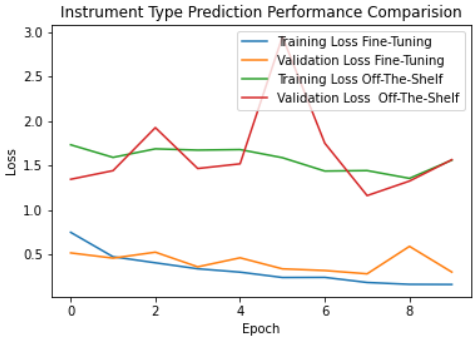


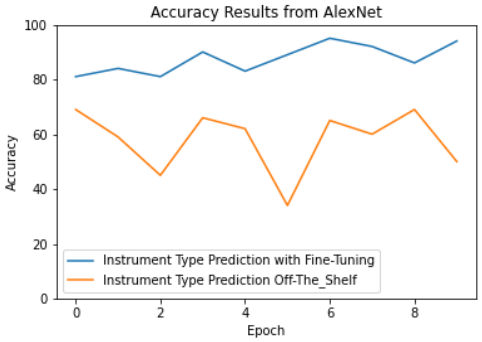


As expected, option 1 worked better than option 2. For option 2, the validation loss plateau at 1.97 with an accuracy of 45 percent, while the validation loss for Option 1 was at 1.6 with an accuracy as high as 62% during training. As seen in the loss and accuracy plots, AlexNet does better with global image standardization than local standardization. AlexNet even with an SGD optimizer gets stuck and cannot learn. This is possibly due to the fact that AlexNet is not good enough at identifying the country made of an instrument based on the shape of the instrument, and instead makes predictions mostly on comparing the color of the instrument to other instruments.

Computational Experiment 3: Off-the-Shelf versus Fine-Tuning

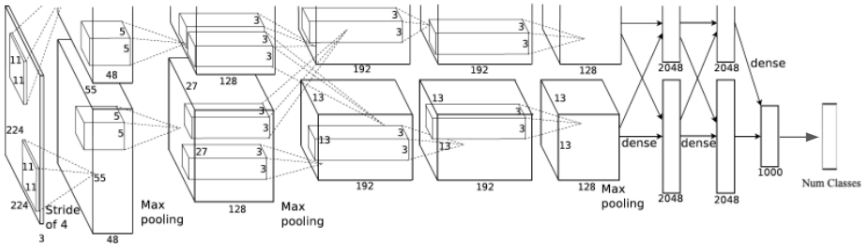
Question:
Will fine-tuning all layers of AlexNet lead to increased performance compared to freezing the weights in earlier layers? Which method will lead to a better model for predicting instrument type and country made?
Option 1:
Only Modify the final Layer of AlexNet.
Option 2:
Allow backpropagation and weight updating throughout all of AlexNet.
Hypothesis:
Option 2 will work slightly better than option 1, but option 2 will have a larger difference between training and validation loss, indicating overfitting.
Results:
AlexNet with Fine-Tuning performed significantly better than Alexnet off-the-shelf.. One reason for Alexnet performing worse without modifying earlier layers is because Alexnet was not trained in music instruments. AlexNet was trained on over a million images from the ImageNet database, but the database does not contain musical instruments. Fine-tuning allowed for new features to be learned that could not be learned if the earlier layers remained frozen.



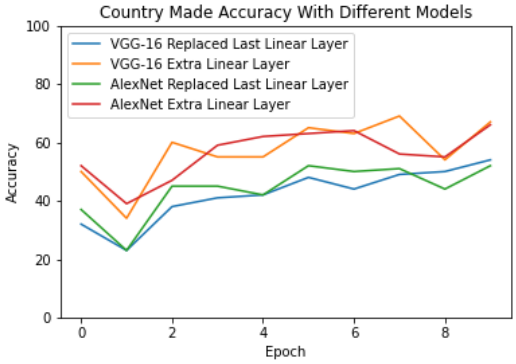


Computational Experiment 4: Testing Different Models
Question:
What other models work better than AlexNet? Would adding an extra layer to AlexNet help?
Methodology:
I will be testing each model for classifying the country made of an instrument, as country prediction has been the lowest performing task.
Option 1:
Use a pre-trained version of VGG-16 with a modified final linear layer.
Option 2:
Use the pre-trained version of VGG-16, but keep the pre-existent last fine-tuned linear layer, and instead add a linear layer to the previous final linear layer.
Option 3:
Use a pre-trained version of AlexNet with a modified final linear layer.

Option 4:
Use the pre-trained version of AlexNet, but keep the pre-existent last fine-tuned linear layer, and instead add a linear layer to the previous final linear layer.

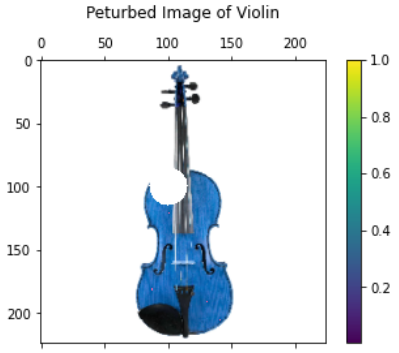


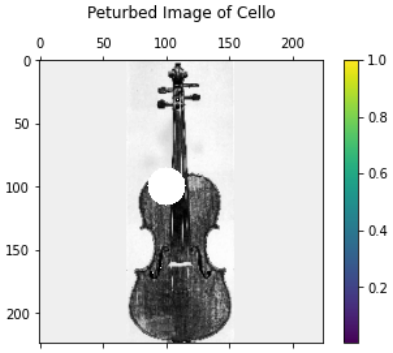
Hypothesis:
VGG-16 will outperform AlexNet, as skip connections and residual connections will allow for better backpropagation through the entire network. The top layers represent problem-specific features, so replacing the final linear layer will achieve higher accuracy of ResNet18.
Results:
Surprisingly, both AlexNet with an extra linear layer and VGG-16 with an extra layer had similar performance. AlexNet and VGG-16 with an extra layer outperformed the models where the final linear layer was replaced. More testing and feature visualization is necessary to understand why keeping the last linear layer increases performance.



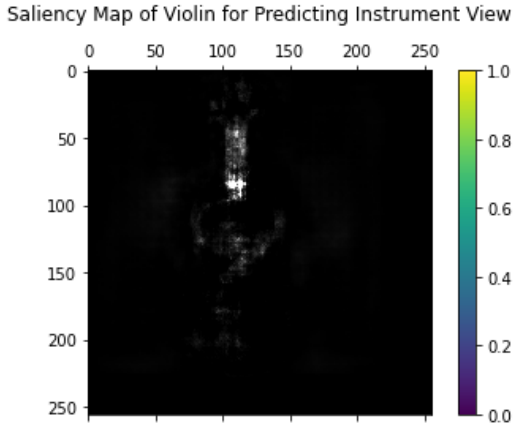
Computational Experiment 5: Creating Class Saliency Map to Detect Important Features

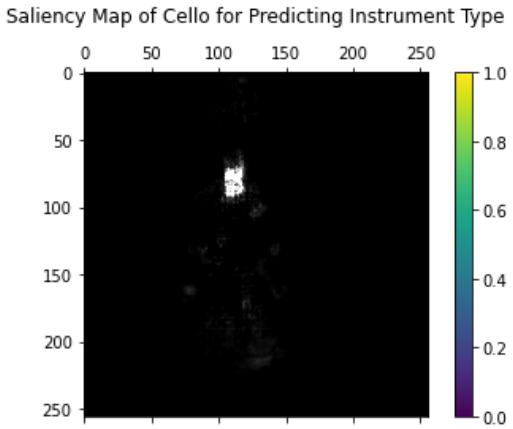
Question:
What parts of a musical instrument are most important for predicting the view and type of instrument?
Methodology:
We decide which part of the image is most important by adding perturbations (white circles or “bubbles”) to an image, and then we attribute the activation of unit i to specific elements of each input. We compute the difference in activations by the following formula:
$\Delta ai(X') = |ai(X) - ai(X')|$. Below is an example of a large perturbation to an input image that decreases the performance of the model. For the saliency map results below, our bubble size is 15 pixels in diameter.





Results:





The central region of the image is important for classification, as all instrument photos are centered in the middle of the image array. Therefore, adding perturbation bubbles to the middle of the image reduces P(correct class | image), therefore causing a larger change in activation. In both saliency maps, the neck of the instrument is the most important part of the image for classification. It is quite understandable that the neck of an instrument is important for classifying the view of an instrument, as a different wood is typically used for the front and back at this location (a light maple for the neck and a dark ebony for the fingerboard).

Challenges Faced:

One of the largest challenges I ran into was running out of RAM and GPU memory while using Google Colab. I solved the issue by creating separate data files that were smaller than the full databases for each classification task. I ensured that the images were only on the GPU when they were being passed into the forward function, rather than having all of the data on the GPU at once. Data normalization was particularly difficult for the project. Several of the photos were clipped off the neck of the instrument, and some of the photos were of poorer quality or black and white.

My models struggled to learn with Adam, and did not see significant performance boosts in later epochs until I switched my optimizer to SGD[3]. Google Colab did not allow for background execution until I figured out a JavaScript console command that tricked Google Colab into thinking that I was clicking the screen when I was away from my computer.

Limitations and Future Work:

I hope to continue this project next semester and try new methods of image processing, feature engineering, and feature extraction. I would also like to use state-of-the-art transformers to predict the price and country made of an instrument. I was previously only passing in one image to make a prediction. Ideally, I could pass in both the front and back of an instrument to substantially increase model performance. One limitation would be to generate multiple mappings of each model instantiation to its respective testing data, allowing accuracy measurements with higher certainty and precision. The loss function could be better optimized and more epochs could have been run. The performance of each model is ultimately bottlenecked by a limited dataset; I hope to gather more data to help increase performance. I hope to continue my work on feature visualization through optimization[4]. With a few lines of code, I was able to make an input optimization model that improved the likelihood of the model predicting the correct class, however the input optimization generated what appeared like random noise to the human eye. I hope to improve my input optimization model to make pixel modifications that reveal important features of the original image.

Conclusion:

AlexNet and other pre-trained deep convolutional neural networks worked extremely well for predicting the type and view of an instrument. However, more work must be done to predict the origin of an instrument with machine learning. One reason for a lack of success on this front is that makers around the world attempt to create instruments that look identical to famous 18th century Italian Violins. A compressed image of the front of an instrument may just not contain enough information to distinguish a French copy of an instrument to the famous Italian instrument that it is trying to copy.

Acknowledgments:

Thank you to Professor Huth and our TA Shailee Jain for advising me through this research. I have learned so much throughout the semester and found an area of research that excites me.

References

1. https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
2. https://paperswithcode.com/sota/image-classification-on-imagenet
3. https://paperswithcode.com/method/sgd
4. https://distill.pub/2017/feature-visualization/